INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION INTERNE

N° 1618

RECOVERY OF EXACT SPARSE REPRESENTATIONS IN THE PRESENCE OF BOUNDED NOISE

JEAN JACQUES FUCHS

ISSN 1166-8687

IRISA

CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Recovery of exact sparse representations in the presence of bounded noise

Jean Jacques Fuchs *

Thème 4 — Simulation et optimisation
de systèmes complexes
Projets SIGMA2

Publication interne n ° 1618 — Avril 2004 — 18 pages

Abstract: The purpose of this contribution is to extend some recent results on sparse representations of signals in redundant bases developed in the noise-free case to the case of noisy observations. The type of questions addressed so far is : given a (n,m)-matrix A with $m > n$ and a vector $b = Ax_o$, find a sufficient condition for b to have an unique sparsest representation. The answer is a bound on the number of non-zero entries of say x_o, that guaranties that it is the unique and sparsest solution of $Ax = b$ when $b = Ax_o$. We consider the case $b = Ax_o + e$ where x_o satisfies the sparsity conditions requested in the noise-free case and e is a vector of additive noise or modeling errors and seek conditions under which x_o can be recovered from b in a sense to be defined.

The conditions we get, relate the noise energy to the signal level as well as to the hyper-parameter of the quadratic program we use to recover the unknown sparsest representation. When the signal-to-noise ratio is large enough all the components of the signal are still present when the noise is deleted, otherwise the smallest components of the signal are themselves erased in a quite rational and predictable way

Key-words: Redundant dictionaries, sparse representations, basis pursuit, global matched filter, quadratic program, non-smooth optimization, mixed ℓ_1-ℓ_2 norm minimization.

(Résumé : tsvp)

soumis comme correspondence à IEEE Transactions on Information Theory

* IRISA /Univ. de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex.—fuchs@irisa.fr

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Solution parcimonieuse de systèmes linéaires sous déterminés en présence de bruit borné.

Résumé : On étend des résultats récents portant sur les solutions parcimonieuses de systèmes linéaires sous déterminés obtenus dans le cas non bruité au cas bruité. Le type de problème traité jusqu'à présent est le suivant: étant donnés une matrice A de dimension (n, m) avec $m > n$ et un vecteur $b = Ax_o$, trouver une condition suffisante pour que b ait une représentation la plus parcimonieuse unique. La réponse est une borne sur le nombre de composantes non-nulles de x_o qui garantit que x_o est la solution la plus parcimonieuse unique de $Ax = Ax_o$. On considère le cas où $b = Ax_o + e$ avec x_o vérifiant la condition de parcimonie et on cherche sous quelles conditions x_o peut être retrouvé dans un sens à préciser à partir de l'observation de b.

On obtient des conditions qui relient l'énergie du bruit à l'amplitude du signal ainsi qu'à un hyper-paramètre du programme quadratique utilisé pour retrouver la représentation la plus parcimonieuse. Quand le rapport signal à bruit est suffisamment grand, toutes les composantes du signal sont encore présentes quand le bruit est effacé, sinon les plus petites composantes du signal sont elles aussi gommées d'une manière prévisible et tout à fait rationnelle.

Mots clés : dictionnaire redondant, représentations parcimonieuses, filtre adapté global, programme quadratique, optimisation de fonction non différentiable.

1 Introduction

Sparse approximation is the problem of finding a representation of a signal as a linear combination of a small number of elements from an over-complete set of vectors or signals often called a dictionary or a redundant basis. Indeed several problems are of interest depending on the context. One may seek the sparsest exact representation of the signal in terms of the elements or the representation of a given complexity that minimizes a certain approximation error or the sparsest representation that yields an approximation error smaller than a specified threshold.

Recently, some theoretical results concerning the first of these questions have been obtained. Given a (n,m)-matrix A with $m > n$ and a vector b that indeed admits an exact sparse representation, say $b = Ax_o$, it has been shown that if the number of non-zero entries in x_o is smaller than a given bound, then x_o is the unique sparsest representation. Since searching for the sparsest representation is a NP hard problem [1] that can only be solved by exhaustive search, one is tempted to replace the true search for the sparsest solution :

$$\min_x \ \|x\|_0 \qquad \text{subject to} \qquad Ax = b \ , \tag{P_0}$$

with $\|x\|_0$ the number of non-zero entries in x, by the easy-to-solve linear program :

$$\min_x \ \|x\|_1 \qquad \text{subject to} \qquad Ax = b \ , \tag{LP}$$

with $\|x\|_1 = \sum_1^m |x_j|$, i.e., to minimize the ℓ_1 norm of x instead of the sparsity itself. The problem is then to determine sufficient conditions for the two criteria to have the same unique solution. This problem has been initiated in [2] and developed since by several other authors, e.g. [3, 4, 5, 8, 6].

We consider an extension of this problem. We assume the signal to have an exact sparse representation that could be recovered in the absence of noise, but that we observe it in additive noise : $b = Ax_o + e$ with e a bounded perturbation vector and seek conditions under which x_o can be recovered partially or completely from the observation of b by solving a convex optimization problem such as a linear or a quadratic program. If one seeks an x that leads to an exact reconstruction of b, it will have generically at least n non-zero components. To get a sparse representation one therefore has to allow for reconstruction errors. The best one can expect is that the optimum of the program and x_o have their non-zero components at the same locations, with the same signs but of course slightly different values. That is we want to recover the support of the true sparse expansion with slightly biased weights, the bias converging to zero as the energy of the noise diminishes.

This problem has received some attention recently [7, 8, 9]. If e is assumed to be Gaussian instead of bounded and no parsimony invoked, it is a difficult detection problem for which numerous ad hoc solutions have been proposed [12] over the last decades in the linear regression literature. It is known as the subset selection or selection of variables problem. In [15] the Gaussian noise case is considered in the present parsimonious representations context and some parts of this contribution can be seen as an application of [15] to the bounded noise case.

This paper is organized as follows. In Section 2, we further specify the problem and introduce the algorithm we shall consider. In Section 3 we present briefly some known results concerning this algorithm that we need in the later sections. The global behavior of this algorithm is investigated in Section 4 both in the case where a sparse representation exists and in the general case of a noisy observation. In Section 5 precise recovery conditions are established when a sparse representation is observed in bounded additive noise. Conclusions are presented in Section 6, they include a summary of the contribution and its practical implications as well as comparisons with other approaches and other results and potential further research directions.

2 Problem formulation

2.1 The noiseless case

Consider a set of m n-dimensional vectors a_j with $m \geq n$ of unit euclidean norm and denote A the (n, m) matrix having these vectors as columns. Any linear combinations b of these m vectors can then be written as $b = Ax$ with x a m-dimensional vector of weights. If $b = Ax_o$, with x_o having just a few non-zero components it may well be the unique and sparsest representation of b. In order to define when this is the case, we introduce the mutual coherence [2]

$$M = \max_{1 \leq i \neq j \leq m} |a_i^T a_j|, \tag{1}$$

of the dictionary whose atoms or components are the columns of A. The smaller M, the less coherent are the components of the dictionary and $M = 0$ if and only if the columns are orthogonal. It is worthwhile to know that there are indeed redundant dictionaries with $m \simeq n^2$ components and mutual coherence $M \simeq \frac{1}{\sqrt{n}}$ [13, 14].

It has been shown in [2, 3, 4, 5] that if

$$\|x_o\|_0 < \frac{1}{2}(1 + \frac{1}{M}) \tag{2}$$

then x_o is the unique sparsest representation of b and that it can be recovered by solving (LP) defined above. In [5], it is further shown that this is also true when (LP) is replaced by

$$\min_x \ \tfrac{1}{2}\|Ax - b\|_2^2 + h\|x\|_1 \ , \quad h > 0, \tag{QP}$$

for sufficiently small h. The problem (QP) is a non-smooth convex optimization problem that can be transform into a quadratic program, hence the denomination. Indeed for $h = 0^+$ (QP) and (LP) are equivalent, but as we shall see below, (QP) is a quite natural way to handle the more general situation where one observes $b = Ax_o + e$ with e a noise or perturbation vector.

2.2 The noisy case

In the sequel, we will investigate the case where one observes $b = Ax_o + e$ with $\|e\|_2 < \epsilon$. We will assume further that $\|x_o\|_0$, satisfies not only (2) but a more stringent condition. Taking a bounded perturbation allows to conduct a deterministic analysis and leads to generic results. In [15], the same parsimonious model is considered but with e a Gaussian noise, the question of the recovery of the true support of x_o is then a complex detection problem and only weak asymptotic results can be expected.

When $b = Ax_o + e$, the optimum of (LP) has generically n non-zero entries and no sparser solution exists generically. One can however expect that for e small enough, the $\|x_o\|_0$ largest, in absolute value, components correspond to the non-zero entries of x_o, while the others are induced by the noise. Applying some pruning to the components of the optimum to null those that are below a threshold could thus lead to the recovery of the true expansion. In the linear regression analysis literature, where m the number of components is generally smaller than n the number of observations, this is known as *elimination of variables* and is based on statistical inference.

We do not pursue this approach since we consider the case where $m > n$ and little can be said about the way the degenerate optimum $x^* = x_o$ of (LP) in the noise-free case evolves as b drifts away from Ax_o. This is due to the fact that for $e = 0$ the optimum of the dual of (LP) is undetermined and it is thus impossible to know a priori which additional components of the optimum of (LP), the primal, will become non-zero. The same remark applies to the following extension of (LP) considered in [16] that allows for reconstruction errors of bounded magnitude in ℓ_∞:

$$\min_x \|x\|_1 \quad \text{subject to :} \quad \|Ax - b\|_\infty \leq B_\infty$$

with $\|x\|_\infty = \max_i |x_i|$. A more promising criterion is $\min \|x\|_1$ subject to $\|Ax - b\|_1 \leq B_1$, but the natural extension to the present ℓ_2-bounded-noise context is indeed:

$$\min_x \|x\|_1 \quad \text{subject to :} \quad \|Ax - b\|_2^2 \leq B_2^2. \tag{3}$$

This is a convex optimization problem that is equivalent to

$$\min_x \ \tfrac{1}{2}\|Ax - b\|_2^2 + h\|x\|_1 \ , \quad h > 0, \tag{QP}$$

for an adequately chosen parameter $h > 0$. This criterion and similar ones have been considered for a while now [17, 10, 11, 7, 8, 9, 6].

Indeed if h in (QP) is taken equal to the inverse of twice the Lagrange multiplier of the constraint in (3) at the optimum, then (3) and (QP) have the same optimum. There is an implicit and unknown relation between B_2 and h. Now (QP) is in turn identical to

$$\min_x \|Ax\|_2^2 \quad \text{subject to} \quad \left\|A^T(Ax - b)\right\|_\infty \leq h \tag{DQP}$$

The two programs (QP) and (DQP) can be transformed into quadratic programs that are equivalent by duality [18, 5]. Of course, (DQP) can be rewritten

$$\min_y \|y\|_2^2 \quad \text{subject to} \quad \left\|A^T(y - b)\right\|_\infty \leq h \tag{DQP'}$$

and the optimum of (DQP) or (QP) deduced from the Lagrange multipliers of the lower dimensional problem (DQP'). In the sequel, we mainly concentrate on (QP) which is the most convenient for our analysis but of course as far as the implementation is concerned any equivalent form can be used.

3 The quadratic programming approach

In this section, we state some properties of the optimum of

$$\min_x \ \frac{1}{2}\|Ax-b\|_2^2 + h\|x\|_1, \quad h>0, \tag{QP}$$

they are established in [5] and we will use them in later sections. The optimum of (QP) will be denoted either x^* or $x(h)$ depending on the context, the notation $x(h)$ being used when the emphasis is on h. We will often need to distinguish between the non-zero components and the zero components of x^*. We denote $\bar{x}^*$ or $\bar{x}(h)$ the reduced dimensional vector built upon the non-zero components of x^*. Similarly $\bar{A}$ denotes the associated columns in A. One then has, e.g., $Ax^* = \bar{A}\bar{x}^*$. We will also use the notation $\bar{\bar{A}}$ for the remaining columns in A and thus decompose A as $A = [\bar{A}\ \bar{\bar{A}}]$.

3.1 Preliminary remarks

Problem (QP) is a non-differentiable convex optimization problem parametrized by a hyper-parameter h.

When $h=0^+$, (QP) is equivalent to (3) with $B_2=0$ which in turn is then identical to (LP). As h increases one can get an intuition about the way the optimum of (QP) evolves by looking at the equivalent criterion (DQP). For $h > \|A^Tb\|_\infty$, $x=0$ is an admissible point in (DQP) that is thus the optimum. One thus expects that, as h increases from 0^+ to $\|A^Tb\|_\infty$, the number of non-zero components in the optimum decreases to become equal to zero.

More precisely, the interval $]0^+,\ \|A^Tb\|_\infty[$ can be divided into sub-intervals characterized by the fact that within each such sub-interval the number of non-zero components of the optimum of (QP) is constant. We will establish below, that while this evolution of the number of non-zero components is not necessarily monotonic at the beginning (for small h) it always is so once the optimum is sparse enough (for large h). This is indeed a valuable feature in a selection of variables context.

Also note that even in the sparse noiseless case, the optimum point $x(h)$ of (QP) achieves a compromise between the two terms of the criterion $\|Ax-b\|_2^2$ and $\|x\|_1$. One therefore has $Ax(h)\neq b$ and thus $\hat{x}(h)\neq x_o$. The best one can expect is that $x(h)$ and x_o have their non-zero components at the same locations, i.e. have the same support, and with the same signs. If this holds we say that *(QP) allows to recover x_o* [5].

In the noisy case, the discrepancy between $x(h)$ and x_o will not only be due to this compromise between the two terms of the criterion but also to the presence of the noise e.

3.2 The optimality conditions

In order to investigate the conditions under which recovery of x_o is possible we need to state and detail the optimality conditions of (QP). For this purpose we introduce the sub-differential of $\|x\|_1$ [19], a set of vectors called the sub-gradients, denoted $\partial\|x\|_1$,

$$\partial\,\|x\|_1 = \{u | u^Tx = \|x\|_1,\ \|u\|_\infty \leq 1\}$$

$$= \{u | u_i = \text{sign}(x_i) \text{ if } x_i \neq 0 \text{ and } |u_i| \leq 1 \text{ otherwise}\}.$$

where $\text{sign}(x_i) = 1$ when $x_i > 0$ and $\text{sign}(x_i) = -1$ when $x_i < 0$. In the scalar case, it amounts to say that the derivative of the function $|x|$, for $x = 0$ is any real in $[-1,\ 1]$. The following result then holds [19].

Theorem 1: The point x^* is a global minimum of (QP) if and only if the vector zero is a sub-gradient of the criterion at x^*:

$$\exists\ u \in \partial \; \|x\|_1 \quad \text{such that} \quad A^T(Ax^* - b) + hu = 0. \qquad \text{(NSC)} \qquad \square$$

To write this necessary and sufficient condition (NSC) in a more usable way, we use the notation $\bar{x}^*$ and $\bar{A}$ defined at the beginning of Section 3, i.e., distinguish between the non-zero components and the zero components of x^*. For the rows in (NSC) associated with the non-zero entries $\bar{x}^*$, the sub-gradient is unique and equal to $\text{sign}(\bar{x}^*)$, for the other rows the sub-gradient takes any value in $[-1,\ 1]$. The necessary and sufficient conditions become

$$\bar{A}^T(b - \bar{A}\bar{x}^*) = h \ \text{sign}(\bar{x}^*), \qquad (\text{NSC}_1)$$

$$|a_j^T(b - \bar{A}\bar{x}^*)| \leq h \ \text{ for } \ a_j \notin \bar{A}. \qquad (\text{NSC}_2)$$

One can further establish that if $\bar{A}$ is full rank, which is always the case if the number of columns in $\bar{A}$ satisfies (2) [8, 6], and if the inequalities in (NSC_2) are strictly satisfied, then the minimum point of (QP) is unique or strict. Hence the following corollary to Theorem 1 [5].

Corollary 1: Sufficient conditions for x^* to be a strict minimum point of (QP) are:

$$1) \qquad \bar{A}^T(b - \bar{A}\bar{x}^*) = h \ \text{sign}(\bar{x}^*), \qquad (4)$$

$$2) \qquad |a_j^T(b - \bar{A}\bar{x}^*)| < h \text{ for } \ a_j \notin \bar{A}, \qquad (5)$$

$$3) \qquad \bar{A} \text{ full rank},$$

with $\bar{x}^*$ and $\bar{A}$ as defined above. $\square$

The first of these relations then leads to an implicit expression of the optimum:

$$\bar{x}(h) = \bar{A}^+ \ b - h \ (\bar{A}^T\bar{A})^{-1} \ \text{sign}(\bar{x}(h)) \qquad (6)$$

with $\bar{A}^+ = (\bar{A}^T\bar{A})^{-1}\bar{A}^T$ a pseudo-inverse of $\bar{A}$.

In (6) we use $x(h)$-notation since the emphasis is on h. This is an implicit relation since $\bar{x}(h)$ is present on both sides. It is valid only for the limited domain around the current value of h for which the signs of all the components of $\bar{x}(h)$ computed according to the expression given in the second member of (6) remain unchanged. Indeed, as h varies, as soon as a component in $\bar{x}(h)$ becomes zero, the decomposition of (NSC) into (NSC_1) and (NSC_2) has to be modified. We take a close look at the behavior of $\bar{x}(h)$ in Section 4.

3.3 The separability and sparsity condition

For later use, we briefly state some results developed in [5]. The idea is to use Corollary 1, to establish that if x_o satisfies (2), then solving (QP) with $b = Ax_o$ and sufficiently small h

allows to recover x_o. Let us decompose, in a similar way as above, x_o into $[\ \bar{x}_o^T\ 0]$ and A into $[\bar{A}_o\ \bar{\bar{A}}_o]$, Theorem 2 in [5] then states.
Theorem 2: The solution x_o of $Ax = b$ with $\bar{A}_o$ and $\bar{x}_o$ as defined above and $\bar{A}_o$ a full rank matrix, can be recovered from the unique optimum point of (QP) if

$$1) \qquad |a_j^T d_o| < 1, \ \ \forall a_j \notin \bar{A}_o \quad \text{with } d_o = \bar{A}_o^{+T}\mathrm{sign}(\bar{x}_o), \tag{7}$$

$$2) \qquad h \in \ ]0,\ h_1[, \ \ \text{the domain in which } \mathrm{sign}\{\bar{x}_o - h\ (\bar{A}_o^T\bar{A}_o)^{-1}\ \mathrm{sign}(\bar{x}_o)\} = \mathrm{sign}(\bar{x}_o) \tag{8}$$

with $\bar{A}_o^+ = (\bar{A}_o^T\bar{A}_o)^{-1}\bar{A}_o^T$, the pseudo-inverse of $\bar{A}_o$. □

Condition (7) is condition (NSC_2) rearranged after substitution of (6), while condition (8) is the translation of the remarks made at the end of the previous section to the present context where h is close to zero. The final step consists in proving that (7) is implied by (2), this is the object of Theorem 3 in [5].
Theorem 3: If the columns in A are normalized to one in ℓ_2 norm, then (2) $\Rightarrow$ (7), i.e.,

$$\|x_o\|_o < \frac{1}{2}(1+\frac{1}{M}) \quad \Rightarrow \quad |a_j^T d_o| < 1, \ \ \forall a_j \notin \bar{A}_o. \qquad \text{where} \qquad d_o = \bar{A}_o^{+T}\mathrm{sign}(\bar{x_o}). \qquad \square \tag{9}$$

Combining then Theorem 2 and 3 leads to the announced recovery result. The proof of Theorem 3 relies on Gershgorin theorem and Neumann's lemma [20]. In the sequel, we will need the following intermediate result [5]:
Lemma: With $\|x_o\|_o = p$ and the notations defined above, one has:

$$|(\bar{A}_o{}^T\bar{A}_o)^{-1}| \leq \alpha I + \beta\mathbf{1}, \qquad \text{with } \alpha = \frac{1}{1+M} \qquad \text{and} \qquad \beta = \frac{M}{(1+M)(1+M-Mp)},$$

where I is the identity matrix, $\mathbf{1}$ a matrix of adequate dimension whose components are all equal to one and where for a matrix B, we denote $|B|$ the matrix with entries $|b_{i,j}|$ and write $B < C$ if $b_{i,j} < c_{i,j}, \forall i,j$. □

4 The behavior of the optimum of (QP)

We are now ready to analyze the behavior of the solution of (QP) as h increases from zero to infinity. We first make some preliminary remarks when no assumption is made on the observation vector b. We later specialize to the case where there exists a true sparse representation satisfying (2) and then pursue with the case where this same representation is observed in additive noise.

4.1 Preliminary remarks

We solve (QP) for a fixed h and analyze what happens to its optimum as h is increased. From Theorem 1 we know that, associated with the optimum x^*, there exists a vector u, a sub-gradient of $\|x\|_1$ at x^*, that satisfies: $A^T(b - Ax^*) = hu$.

As explained at the beginning of Section 3, we decompose x^* into $\bar{x}^*$ and $\bar{\bar{x}}^* = 0$, and accordingly the matrix A and the vector u. Replacing $\bar{u}$ by $\text{sign}(\bar{x}^*)$, the preceding relation becomes

$$\bar{A}^T(b - \bar{A}\bar{x}^*) = h\ \text{sign}(\bar{x}^*)$$
$$\bar{\bar{A}}^T(b - \bar{A}\bar{x}^*) = h\bar{\bar{u}}.$$

We are interested in how $\bar{x}^*$ and $\bar{\bar{u}}$ evolve as h increases because the present decomposition is only valid as long as no component in $\bar{x}^*$ becomes zero and no component in $\bar{\bar{u}}$ becomes equal to ± 1. From the first of these two relations we deduce $\bar{x}^*$ which we denote $\bar{x}(h)$ to emphasize the dependency on h. We introduce this expression into the second relation to finally obtain after some rearrangements

$$\begin{aligned} \bar{x}(h) &= \bar{A}^+\ b - h\ (\bar{A}^T\bar{A})^{-1}\ \text{sign}(\bar{x}(h)) \\ \bar{\bar{u}}(h) &= \frac{1}{h}\bar{\bar{A}}^T b^\perp + \bar{\bar{A}}^T\bar{A}^{+T}\text{sign}(\bar{x}(h)) \end{aligned} \tag{10}$$

where $b^\perp = (I - \bar{A}\bar{A}^{+T})b$ is the projection of b on the subspace orthogonal to the range of $\bar{A}$. Since $\|\bar{x}(h)\|_1 = \text{sign}(\bar{x}(h))^T\bar{x}(h)$, it follows from the first of these two relations that for small δh

$$\|\bar{x}(h+\delta h)\|_1 = \|\bar{x}(h)\|_1 - \delta h\ \text{sign}(\bar{x}(h))^T(\bar{A}^T\bar{A})^{-1}\ \text{sign}(\bar{x}(h)),$$

which shows that the ℓ_1-norm of $\bar{x}(h)$ decreases as h increases. If in the second relation, we introduce, as in Theorem 2, the vector $d = \bar{A}^{+T}\text{sign}(\bar{x}(h))$, it becomes

$$\bar{\bar{u}}(h) = \frac{1}{h}\bar{\bar{A}}^T b^\perp + \bar{\bar{A}}^T d. \tag{11}$$

One can then check that

$$\|\bar{\bar{u}}(h)\|_\infty = \|\frac{1}{h}\bar{\bar{A}}^T b^\perp + \bar{\bar{A}}^T d\|_\infty \leq 1 \quad \text{and} \quad \|\bar{\bar{A}}^T d\|_\infty < 1 \quad \Rightarrow \quad \|\bar{\bar{u}}(h+\delta h)\|_\infty < 1,\ \forall\, \delta h > 0 \tag{12}$$

where the verification can de done for each $a_j \in \bar{\bar{A}}$. The relations (10, 11, 12) will be used below to investigate the evolution of the optimum $x(h)$ of (QP) as h increases from 0^+ to infinity.

Remark: The implication (12) brings into light the importance in our context of the vector d. It was introduced in Theorem 2 and somehow decides whether a representation is recoverable or not. It appears that a representation is essentially defined by $\bar{A}$, the columns it uses and $\text{sign}(\bar{x})$, the signs of the corresponding weights. This is also precisely what is needed to build $d = \bar{A}^{+T}\text{sign}(\bar{x})$. One associates with this single vector d [17, 5] two parallel *separating* hyperplanes $H_\pm$, defined by: $H_\pm = \{a|a^T d = \pm 1\}$ and Theorem 2, then says that (QP) allows to recover the representation if the columns of A not in $\bar{A}$ (those in $\bar{\bar{A}}$) lie strictly in between these two hyperplanes (7). If the emphasis is put on sparse representations, it follows that a representation is recoverable if the sparsity condition (2) implies the separability condition (7), see Theorem 3.

4.2 The sparse noiseless case

Let $b = Ax_o$ with x_o satisfying the parsimony condition (2). Solving (QP) for $h = 0^+$ yields then an optimum $x(0^+)$ that allows to recover x_o. This is established in [5] and follows from Theorem 2 and 3. For h small enough, relations (10) and (11) become

$$\begin{aligned} \bar{x}_o(h) &= \bar{A}_o^+ b - h\ (\bar{A}_o^T \bar{A}_o)^{-1} \mathrm{sign}(\bar{x}_o(h)) = \bar{x}_o - h\ (\bar{A}_o^T \bar{A}_o)^{-1} \mathrm{sign}(\bar{x}_o) \\ \bar{\bar{u}}_o(h) &= \frac{1}{h} \bar{\bar{A}}_o^T b_o^\perp + \bar{\bar{A}}_o^T d_o = \bar{\bar{A}}_o^T d_o \end{aligned} \tag{13}$$

with $d_o = \bar{A}_o^{+T} \mathrm{sign}(\bar{x}_o)$. Note that $b_o^\perp = 0$ follows from $\bar{A}_o^+ b = \bar{x}_o$. To distinguish this initial decomposition, that is associated with x_o, from those that will be performed below, we added a subscript "$_o$".

The vector $\bar{\bar{u}}_o(h)$ is thus independent of h for small h. This is not the case for the components in $\bar{x}_o(h)$ in (13). As h increases, $\|\bar{x}_o(h)\|_1$ decreases and a first component of $\bar{x}_o(h)$ will become zero. As in Theorem 2, we denote h_1 the corresponding value of h. At this point, one re-decomposes the optimum into $[\bar{x}_1(h)^T\ 0]^T$, A into $A = [\bar{A}_1\ \bar{\bar{A}}_1]$ and u accordingly. One then has, for h slightly larger than h_1,

$$\begin{aligned} \bar{x}_1(h) &= \bar{A}_1^+ b - h\ (\bar{A}_1^T \bar{A}_1)^{-1} \mathrm{sign}(\bar{x}_1(h)) \\ \bar{\bar{u}}_1(h) &= \frac{1}{h} \bar{\bar{A}}_1^T b_1^\perp + \bar{\bar{A}}_1^T d_1 \end{aligned} \tag{14}$$

with $b_1^\perp$ the projection of b on the subspace orthogonal to the range of $\bar{A}_1$ and d_1 defined similarly to d_o above. As h increases beyond h_1 both the components in $\bar{\bar{u}}_1(h)$ (14) and those in $\bar{x}_1(h)$ vary and one has to check which event happens first at say $h = h_2$, a component in $\bar{\bar{u}}_1(h)$ becoming equal to ± 1 or a component in $\bar{x}_1(h)$ becoming zero.

Since $\bar{x}_1(h)$ satisfies (2), it follows from Theorem 3 that $\|\bar{\bar{A}}_1^T d_1\|_\infty < 1$ and hence from implication (12) that $\bar{\bar{u}}_1(h)$ will remain strictly inside $[-1, 1]$. At $h = h_2$, a component of $\bar{x}_1(h)$ will thus become zero and this will further diminish the number of non-zero components in $x(h)$. The same reasoning can then be re-conducted until $x(h) = 0$ for $h \geq \|A^T b\|_\infty$. We have thus established.

Proposition 1: If x_o satisfies the parsimony condition (2), (QP) allows to recover it for h sufficiently small, i.e. for $h \in]0^+, h_1[$. If h is increased beyond h_1 the number of non-zero components in the optimum $x(h)$ of (QP) decreases monotonically. □

One can further check that $x(h)$ is a continuous function of h and that $\|x(h)\|_1$ is monotonically decreasing toward zero. Note that the implication in (12) is quite generally applicable and actually holds for any b and any h as soon as the corresponding optimum $x(h)$ satisfies (2).

4.2.1 The general case

We now consider an arbitrary b, this includes of course the case where $b = Ax_o + e$ with $\|e\|_2 \leq \epsilon$ and x_o satisfies the parsimony condition (2). Solving (QP) for $h = 0^+$ is equivalent to solving the linear program (LP) whose solution is generically obtained at a basic feasible solution [19]. This means that $x(0^+)$ has n non-zero components and that in the associated partition of $A = [\bar{A}_o \; \bar{\bar{A}}_o]$, $\bar{A}_o$ is invertible. With the notations of the previous section, one has :

$$\begin{aligned} \bar{x}_o(h) &= \bar{A}_o^{-1} b - h \; (\bar{A}_o^T \bar{A}_o)^{-1} \mathrm{sign}(\bar{x}_o(0^+)) \\ \bar{\bar{u}}_o(h) &= \frac{1}{h} \bar{\bar{A}}_o^T b_o^\perp + \bar{\bar{A}}_o^T d_o = \bar{\bar{A}}_o^T d_o \end{aligned} \tag{15}$$

Note that now $b_o^\perp = 0$ because the range of $\bar{A}_o$ is the whole space. For h sufficiently small $\bar{\bar{u}}_o(h)$ is thus again independent of h and as h increases a component in $\bar{x}_o(h)$ will become zero at $h = h_1$ since $\|\bar{x}_o(h)\|_1$ decreases. Assume it is x_{on} the last component in $\bar{x}_o(h)$ that is associated with column a_n of $\bar{A}_o$. One then re-decomposes the optimum into $[\bar{x}_1(h)^T \; 0]^T$, A into $A = [\bar{A}_1 \; \bar{\bar{A}}_1]$ and u accordingly. Using the formula giving the partitioned inverse of $\bar{A}_o$ written as $\bar{A}_o = [\bar{A}_1 \; a_n]$, the last component x_{on} of $\bar{x}_o(h)$ in (15), that becomes zero for $h = h_1$, is equal to:

$$x_{on}(h) = \frac{1}{\|a_{n1}^\perp\|^2} \; \{a_n^T b_1^\perp + h a_n^T d_1 - h \mathrm{sign}(x_{on})\} \tag{16}$$

where $b_1^\perp$ and $a_{n1}^\perp$ are the projections of b and a_n on the subspace orthogonal to the range of $\bar{A}_1$ and d_1 is defined as above. For $h = h_1$, one has the following two expressions for $\bar{\bar{u}}_1(h_1)$:

$$\bar{\bar{u}}_1(h_1) = \begin{bmatrix} \mathrm{sign}(x_{on}) \\ \bar{\bar{A}}_o^T d_o \end{bmatrix} = \frac{1}{h_1} \; \bar{\bar{A}}_1^T b_1^\perp + \bar{\bar{A}}_1^T d_1$$

As h increases beyond h_1 both $\bar{\bar{u}}_1(h) = \frac{1}{h} \; \bar{\bar{A}}_1^T b_1^\perp + \bar{\bar{A}}_1^T d_1$ and $\bar{x}_1(h) = \bar{A}_1^+ b - h(\bar{A}_1^T \bar{A}_1)^{-1} \mathrm{sign}(\bar{x}_1(h_1))$ are varying and it is now possible that a component in $\bar{\bar{u}}_1(h)$ becomes equal to ± 1 before a further component of $x(h)$ becomes zero. Using (16), one can verify that it will not be the first component of $\bar{\bar{u}}_1(h)$, the one that just entered $\bar{\bar{u}}_1(h)$. Let us assume it is the component of $\bar{\bar{u}}_1$ associated with, say, the column a_2 of $\bar{\bar{A}}_1$ that becomes equal to, say, 1 for $h = h_2$. This means that: $\frac{1}{h_2} a_2^T b_1^\perp + a_2^T d_1 = 1$ and for later use note that since this quantity would become greater than $+1$ for larger h, one has that $a_2^T b_1^\perp < 0$ and $a_2^T d_1 > 1$.

At $h = h_2$ one re-decomposes A and $\bar{A}_2 = [\; \bar{A}_1 \; a_2]$. The new non-zero components in $x(h)$ included in $\bar{x}_2(h)$ say $x_{22}(h)$ has then the following expression, using the partitioned form of $\bar{A}_2$,

$$x_{22}(h) = \frac{1}{\|a_2^\perp\|^2} \; \{a_2^T b_1^\perp + h a_2^T d_1 - h \mathrm{sign}(x_{22}(h))\} = \frac{1}{\|a_2^\perp\|^2} \; \{a_2^T b_1^\perp + h a_2^T d_1 - h\}$$

it is equal to zero for $h = h_2$ by definition of h_2, and, as follows from $a_2^T b_1^\perp < 0$ and $a_2^T d_1 > 1$, becomes strictly positive as h increases. It will thus certainly not become zero during the next interval $h \in]h_2, \; h_3]$.

This step completes the analysis of the behavior of the optimum $x(h)$ of (QP) as h increases. In summary, one can say that, as h increases, $\|x(h)\|_1$ decreases steadily, but this does not hold for $\|x(h)\|_0$, the number of non-zero components in $x(h)$. This number which never exceeds n, can sometimes increase. This last event however cannot happen once $\|x(h)\|_0$ satisfies (2) as we have seen in section 4.2. Note that since (2) is only a sufficient condition, this nice rational behavior can start for larger values of $\|x(h)\|_0$.

5 Recovery properties of (QP)

In the previous section we investigated the general behavior of the optimum $x(h)$ of (QP), we will now analyze more precisely its properties. We establish that

Theorem 4: Let b be equal to $b = Ax_o + e$ with $\|e\|_2 \leq \epsilon$ and $p = \|x_o\|_0 \leq \frac{\beta}{M}$ with $\beta \leq \frac{1}{2}$. Then , if one solves (QP) with $h > (1+\frac{\beta}{1+M-2\beta})\epsilon$ the support of the optimum $x(h)$ is either identical or contained in the support of x_o.
And if one solves (QP) with $h = (1+\frac{\beta}{1+M-2\beta})\epsilon^+$ and if $x_m > (\frac{1}{1+M-\beta}+\frac{1}{1+M-2\beta})\epsilon$, with x_m the smallest in absolute value non-zero component of x_o, then one recovers the true representation, i.e., $x(h)$ and x_o have their non-zero components at the same locations and with the same signs. □

Proof:
Solving (QP), one knows from Theorem and Corollary 1 that $x(h)$ is the optimum if and only if $\exists u \in \partial\|x\|_1$ such that $A^T(b - Ax(h)) = hu$, (NSC). We decompose x_o into $x_o^T = [\bar{x}_o^T\ 0]$ with $\bar{x}_o$ the p-dimensional vector of the non-zero components of x_o and proceed similarly with the matrix A, the optimum $x(h)$ and the vector u. One then has, for instance, $b = Ax_o + e = \bar{A}_o\bar{x}_o + e$. Note that this decomposition is driven by x_o and not $x(h)$ and that e.g. $\bar{u}$ is not necessarily equal to $\mathrm{sign}(\bar{x}(h))$.

To establish the first part of the theorem, we need to show that $\|\bar{\bar{u}}\|_\infty < 1$ since this implies that $\bar{\bar{x}}(h) = 0$ as follows from the definition of $\partial\|x\|_1$.

Decomposing the relation (NSC) into two parts, the first $\bar{A}_o^T(b - \bar{A}_o\bar{x}(h)) = h\bar{u}$ leads to:

$$\bar{x}(h) = \bar{A}_o^+ b - h(\bar{A}_o^T\bar{A}_o)^{-1}\bar{u} = \bar{x}_o + \bar{A}_o^+ e - h(\bar{A}_o^T\bar{A}_o)^{-1}\bar{u}$$

with some components potentially equal to zero. The second part of (NSC) $\bar{\bar{A}}_o^T(b - \bar{A}_o\bar{x}(h)) = h\bar{\bar{u}}$ becomes after substitution of $\bar{x}(h)$:

$$\bar{\bar{A}}_o^T(e^\perp + h\delta_o) = h\bar{\bar{u}} \qquad \text{with} \qquad \delta_o = \bar{A}_o^{+T}\bar{u}$$

where $e^\perp = (I - \bar{A}_o\bar{A}_o^+)e$. To establish $\|\bar{\bar{u}}\|_\infty < 1$ amounts thus to prove that $|a_j^T(e^\perp + h\delta_o)| < h,\ \forall a_j \in \bar{\bar{A}}_o$. Remember that by definition of u one always has $\|u\|_\infty \leq 1$.

Now $|a_j^T e^\perp| \leq \epsilon$ since $\|e\|_2 \leq \epsilon$ and $\|a_j\|_2 = 1$. Using the Lemma of Section 3.3, the definition of the mutual coherence M and the fact that $\|\bar{u}\|_\infty \leq 1$ one has:

$$|a_j^T\delta_o| \qquad \leq |a_j^T\bar{A}_o||(\bar{A}_o{}^T\bar{A}_o)^{-1}||\bar{u}|$$

$$\begin{aligned}&\leq M\mathbf{1}^T|(\bar{A}_o{}^T\bar{A}_o)^{-1}|\mathbf{1}\\&\leq M\ (\alpha p+\beta p^2)\\&\leq \frac{Mp}{1+M-Mp}\end{aligned}$$

On the other hand: $|a_j^T(e^\perp + h\delta_o)| \leq |a_j^T e^\perp| + h|a_j^T\delta_o| \leq \epsilon + h\frac{Mp}{1+M-Mp}$. Thus $h > \epsilon + h\frac{Mp}{1+M-Mp}$ implies $|a_j^T(e^\perp + h\delta_o)| < h$ and this holds precisely if the announced bound for h is satisfied. This completes the proof of the first part of the theorem.

For the proof of the second part of the theorem, we need to prove as above that $\|\bar{u}\|_\infty < 1$ and that under the additional condition on x_m no component of $\bar{x}(h)$ is zero and $\text{sign}(\bar{x}(h)) = \text{sign}(\bar{x}_o)$. To establish that $\|\bar{u}\|_\infty < 1$, one proceeds exactly as above, the proof goes through without modification. It remains to establish that $\text{sign}(\bar{x}(h)) = \text{sign}(\bar{x}_o)$, we start from

$$\bar{x}(h) = \bar{A}_o^+\ b - h\ (\bar{A}_o^T\bar{A}_o)^{-1}\ \bar{u} = \bar{x}_o + \bar{A}_o^+\ e - h\ (\bar{A}_o^T\bar{A}_o)^{-1}\ \bar{u}.$$

Using the Lemma of Section 3.3, one has

$$\begin{aligned}&\|\bar{A}_o^+\ e\|_\infty = \|(\bar{A}_o^T\bar{A}_o)^{-1}\ \bar{A}_o^Te\|_\infty \leq \epsilon\|\ |(\bar{A}_o{}^T\bar{A}_o)^{-1}|\mathbf{1}\ \|_\infty \leq \frac{\epsilon}{1+M-Mp} \leq \frac{\epsilon}{1+M-\beta}\\&\|(\bar{A}_o^T\bar{A}_o)^{-1}\ \text{sign}(\bar{x}_o)\|_\infty \leq \|\ |(\bar{A}_o{}^T\bar{A}_o)^{-1}|\mathbf{1}\ \|_\infty \leq \frac{1}{1+M-Mp} \leq \frac{1}{1+M-\beta}\end{aligned}$$

Combining these results, for $\text{sign}(\bar{x}(h)) = \text{sign}(\bar{x}_o)$ to hold, one must have that $x_m > \frac{\epsilon+h}{1+M-\beta}$ which for $h = (\frac{1+M-\beta}{1+M-2\beta})\epsilon^+$ leads to the announced bound for x_m. □

6 Conclusions

6.1 Summary

In practice, one can thus solve (QP) for increasing h. As soon as $\|x(h)\|_0 < \frac{1}{2}(1+\frac{1}{M})$, the optimum of the algorithm has a quite rational behavior with the number of non-zero components further decreasing when h is further increased. If one assumes to know ϵ, one can further increase h to finally reach $\tilde{h}$ the smallest value of h for which it holds that

$$\tilde{\beta} < \frac{1}{2} \quad \text{and} \quad \tilde{h} \geq (1 + \frac{\tilde{\beta}}{1+M-2\tilde{\beta}})\epsilon$$

with $\tilde{\beta} = M\|x(\tilde{h})\|_0$. Such an $\tilde{h}$ always exist since $\tilde{\beta}$ and the right side of the second inequality decrease as h increases. For $h \geq \tilde{h}$, all the retained components belong necessarily to the true sparsest representation.

6.2 Comparison to other methods and results

Iterative approaches to finding sparse representations have been around for far longer than global approaches as the one presented here. Recently it has been shown that though being fundamentally suboptimal, they possess quite interesting properties [8]. This is especially true for the approach called Orthogonal Greedy algorithm (OGA) in [7] or Orthogonal Matching Pursuit in [8] which was introduced in the early nineties in a effort to improve upon the basic Matching Pursuit (MP) algorithm [10].

In order to describe these algorithms, we denote $r^{(k)}$ the reconstruction error or the residual at step k with initial residual $r^{(0)} = b$. (MP) is then an iterative method which at step k picks the column-vector of the A matrix with index $j_k = \arg\max_j |a_j^T r^{(k-1)}|$ that has maximal correlation with $r^{(k-1)}$ and assigns to it a weight $\chi_{j_k} = a_{j_k}^T r^{(k-1)} = \arg\min_\chi \|r^{(k-1)} - \chi\ a_{j_k}\|_2^2$ that minimizes the energy of the new residual vector $r^{(k)} = r^{(k-1)} - \chi_{j_k}\ a_{j_k}$. This basic step is repeated until a stopping criterion is satisfied. The following rules are common: stop after a specified number of atoms have been selected, stop when the maximal correlation is below a threshold or stop when the residual norm is smaller than some fixed threshold.

In (OGA), the difference with (MP) lies in the re-evaluation of the weights. At the end of step k when the new entering column a_{j_k} has been selected as in (MP), all the weights now denoted $\{\chi_{j_i}^{(k)},\ i = 1,..,k\}$ are recomputed to minimize the residual norm

$$\min_{\chi_{j_i},\ i=1,..,k} \ \|b - \sum_{i=1}^{k} \chi_{j_i}\ a_{j_i}\|_2^2$$

This least squares fit corresponds to an orthogonal projection of b on the subspace spanned by the k atoms in $\{a_{j_i},\ i = 1,..,k\}$, hence the name of the procedure. If we denote $\bar{A}_{(k)}$ the matrix whose columns are the k selected atoms then the optimal weights are given by $\bar{A}_{(k)}^{+} b$, with $\bar{A}_{(k)}^{+} = (\bar{A}_{(k)}^T \bar{A}_{(k)})^{-1} \bar{A}_{(k)}^T$.

It is interesting to notice that (QP) in some sense achieves a goal that is pretty similar to the one of (OGA) when the stopping rule is set on the maximal acceptable correlation. Indeed looking at the dual (DQP) of (QP) [18, 5]:

$$\min_x \|Ax\|_2^2 \quad \text{subject to} \quad \|A^T(Ax - b)\|_\infty \leq h \qquad \text{(DQP)}$$

one realizes that the constraint in this problem corresponds precisely in seeking a representation x that is such that the maximal correlation between the residual or reconstruction error $b - Ax$ and any column in A is smaller than h in absolute value. This means that (QP) performs, according to a minimal energy criterion, in a global way [18] and in a single, possibly computationally intensive, step what (OGA) achieves iteratively and sub-optimally trough a much computationally cheaper relaxation procedure.

There is however another difference that makes the two approaches distinct. At the optimum x^* of (QP) the correlations of the retained atoms with the final residual are set to $\pm h$ while in (OGA) they are set to zero. The non-zero components $\bar{x}^*$ of the optimum x^* of (QP) satisfy (6):

$$\bar{x}^* = \bar{A}^{+}\ b - h\ (\bar{A}^T \bar{A})^{-1}\ \text{sign}(\bar{x}^*)$$

while the weights of the atoms retained by (OGA) are $\bar{A}^+_{(k)}b$. There is thus no straight connection between both methods, because even though the second term in $\bar{x}^*$, which can be considered as a bias term induced by the presence of the regularization term in (QP), can be remove a posteriori to keep just $\bar{A}^+\, b$, one does not know which criterion, if any, this new point optimizes.

To our knowledge, the scenario we consider in this article has first been proposed in [7], where among other results it is established that if (OGA) is used with a stopping criterion using the threshold ϵ on the ℓ_2-norm of the residual, and provided the smallest in absolute value non-zero component of the true representation is greater than a given bound, then (OGA) recovers the true support. This is also the type of results we propose in the second part of Theorem 4 for (QP). Though it is impossible to compare the two results since they use different bounds, one can observe that the threshold on x_m we get is slightly smaller.

In [7], a further result is proposed that corresponds to the first part of Theorem 4 and holds for (QP) in its equivalent form (3). Again it seems impossible to compare the results since there is no analytical link between B_2 in (3) and h in (QP).

6.3 Extensions

In [15], a similar model is considered in conjunction with (QP) but with the assumption that the noise e is Gaussian instead of ℓ_2 bounded. The recovery of the exact sparse representation is then a difficult detection problem and only weak asymptotic, in the signal to noise ratio, results can be expected. The analysis, which has some similarities with the one where more general scenarios are considered. Indeed in a detection-estimation context, many practical situations correspond to the case where the true model for the observed vector b is

$$b = \sum_{i=1}^{p} \alpha_i a(\tau_i) + e,$$

with $a(\tau)$ a known family of vectors parametrized by a scalar τ, α_i the scalar weights and e some additive noise. In order to apply the previous setting to this scenario where one wants to detect the number p of components and estimate theirs weights α_i and parameters τ_i a natural way to proceed is to uniformly discretize the values of τ over its compact domain to get the m columns $a_j = a(\tau)|_{\tau=t_j}$ of the A matrix and to apply for instance (QP). Since the true τ_i will generically not belong to the sampling points used to build the columns in A, there is no exact sparse representation of b even if p the true number of components is small. Hence the necessity to tune the hyper-parameter h in (QP) not only in order to erase the noise e but also to take care of the approximation errors that depend upon the discretization step. In the case of a single noiseless component, when $b = \alpha_1 a(\tau_1)$, the weights of x^* the optimum of (QP) can be shown to be close to samples of a minimum L_1-norm interpolating function [21]. A more general scenario is considered in [18] where the issue is super-resolution. Heuristic rules are presented that indicate how both the discretization step and the parameter h in (QP) should be chosen as a function of the signal to noise ratio and the expected performances evaluated by means of the Cramer-Rao bounds.

References

[1] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, 24, 21, 227-234, April 1995.

[2] D.L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. on I.T.*, 47, 11, 2845-2862, Nov. 2001.

[3] M. Elad an A.M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. on I.T.*, 48, 9, 2558-2567, Sept. 2002.

[4] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. on I.T.*, 49, 12, 3320-3325, Dec. 2003.

[5] J.J. Fuchs, "More on sparse representations in arbitrary bases," *13th Ifac-Ifors symposium on Identification and system Parameter Estimation*, pp. 1357-1362, Rotterdam, August 2003, to appear in *IEEE Trans. on I.T.*, June 2004.

[6] D. Malioutov, M. Cetin and A.S. Willsky "Optimal sparse representations in general overcomplete bases," to be presented at *IEEE ICASSP*,Montreal, May 2004.

[7] D.L. Donoho, M. Elad and V. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," submitted to *IEEE Trans. on I.T.*, Feb. 2004.

[8] J.A. Tropp, "Greed is good: Algorithmic Results for Sparse Approximations," ICES Report 0304, UT-Austin, Feb. 2003, to appear in *IEEE Trans. on I.T.*

[9] J.A. Tropp, "Just relax: Convex Programming Methods for Subset Selection and Sparse approximation," ICES Report 0404,UT-Austin, Feb. 2004.

[10] S. Chen, D. Donoho and M. Saunders, "Atomic Decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, Jan. 1999.

[11] J.J. Fuchs, "Multipath time-delay detection and estimation," *IEEE Trans. Signal Processing*, vol. 47, no. 1, pp. 237–243, Jan. 1999.

[12] N.R. Draper and H. Smith. Applied Regression. *John Wiley and Sons*, 1966.

[13] T. Strohmer and R. Heath, "Grassmanian frames with applications to coding and communications," to appear in *Appl. Comp. Harm. Analysis.*

[14] A. Gilbert, S. Muthukrishnan and M.J. Strauss, "Approximation of functions over redundant dictionaries using coherence," *14th ACM-SIAM Symposium on Discrete Algorithms*, (SODA'03), Jan. 2003.

[15] J.J. Fuchs, "Recovery of exact sparse representations in the presence of noise," to be presented at *IEEE ICASSP*,Montreal, May 2004.

[16] J.J. Fuchs, "Extension of the Pisarenko method to sparse linear arrays," *IEEE Trans. Signal Processing*, vol. 45, pp. 2413–2421, Oct. 1997.

[17] J.J. Fuchs, "Detection and estimation of superimposed signals," *IEEE ICASSP*, vol. III, pp. 1649–1652, Seattle, U.S.A., 1998.

[18] J.J. Fuchs, "On the application of the global matched filter to DOA estimation with uniform circular arrays," *IEEE-T-SP*, vol. 49, p. 702–709, April 2001. *IEEE Signal Processing Society 2003 Best Paper Award.*

[19] R. Fletcher. Practical methods of optimization. *John Wiley and Sons*, 1987.

[20] A.S. Householder. *The theory of matrices in numerical analysis.* New York, N.Y.: Blaisdell, 1964.

[21] J.J. Fuchs and B. Delyon, "Minimum L_1-norm reconstruction function for oversampled signals: Application to time-delay estimation," *IEEE Trans. Inform. Theory*, vol. 46, no. 4, 1666–1673, Jul. 2000.

Contents